DOCUMENT RESUME

ED 246 108 TM 840 387

AUTHOR Korpi, Meg; Haertel, Edward
TITLE Locating Reading Test Items in Multidimensional Space: An Alternative Analysis of Test Structure.
PUB DATE Apr 84
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Correlation; *Factor Structure; Intermediate Grades; Item Analysis; *Multidimensional Scaling; *Reading Tests; Statistical Analysis; Test Items; *Test Validity
IDENTIFIERS Data Interpretation; Dichotomous Variables; Metropolitan Achievement Tests

ABSTRACT

The purpose of this paper is to further the cause of clarifying construct interpretations of tests, by proposing that non-metric multidimensional scaling may be more useful than factor analysis or other latent structure models for investigating the internal structure of tests. It also suggests that typical problems associated with scaling dichotomous variables can be avoided by using tetrachoric correlations as input to the multidimensional scaling. Finally, it demonstrates the utility of the suggested procedures by applying them to actual data from the Reading subtest of the Metropolitan Achievement Tests. (BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED246108

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Locating Reading Test Items in Multidimensional Space: An alternative analysis of test structure

Meg Korpi and Edward Haertel
Stanford University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

M. Korpi

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the Annual Meeting of the
American Educational Research Association, New Orleans, LA, April, 1984

TM 840 387

Locating Reading Test Items in Multidimensional Space: An alternative analysis of test structure

Test scores play a major role in the public's perception of how well schools educate students. Recent surveys of test use in schools have indicated that a prime use of standardized achievement test data is for accountability, that is, reporting to the public (e.g., see case studies in Hathaway, 1983, and Herman & Dorr-Bremme, 1983). In the media, schools are continually judged according to student test scores. The recent widely publicized study of education in the United States by the prestigious National Commission on Excellence in Education (1983) listed thirteen indicators of "a nations at risk." Ten of those indicators involve test results. Yet, with all this concern for how well schools are doing, little attention has been paid to how well the tests are doing. Are they fair criteria by which to judge education? What do they really tell us?

We make the preceding point to emphasize the importance of construct validity in test use. Construct validity is more than an intellectual issue for test developers; when tests are interpreted and used by the public, it becomes a public issue. Cronbach (1980) has stated that test interpreters are responsible for validating a test for a particular use, but that measurement professionals must provide information that helps clarify what tests measure, so that interpreters of tests might use them more wisely. Our purpose in this paper is to further the cause of clarifying construct interpretations of tests, by proposing an alternative technique for analyzing underlying test structure. Factor analysis is the method most commonly used for this purpose. However, a variety of problems are associated with it. We propose that non-metric multidimensional scaling (MDS) may be more useful than factor analysis or other latent structure models for investigating the internal structure of tests. We then demonstrate the utility of this technique by applying it to data from a widely used standardized test of reading comprehension.

Background

Comparison of MDS and Factor Analysis

Multidimensional scaling is a data representation technique for showing the relationships between objects by locating them as points in a continuous space (Kruskal & Wish, 1978). In the spatial representation, similar objects appear closer together and dissimilar ones appear farther apart. The technique has been used for a variety of psychological purposes, and may be applied to any set of objects for which measures of similarity (or dissimilarity) are available. The method requires few assumptions--basically, that the similarity measures can be represented as Euclidean distances. Unlike factor analysis, it requires no special assumptions about the underlying processes giving rise to the similarity data.

The most common procedure for examining the internal structure of single tests, or the structure of batteries of tests, has been factor analysis. Unlike multidimensional scaling, factor analysis sets forth and fits an explicit model to a matrix of covariances or correlations. It makes the strong assumption that each observed variable is a weighted sum of some small number of common, unobservable variables called factors. In addition to the

common factors, there is a different unique factor for each observed variable. These unique factors are assumed to be uncorrelated with the common factors and with each other. As a consequence of these assumptions, the unique factors contribute to their respective variables' variances, but not to their covariances. The covariances are entirely due to the common factors.

Factor analysis requires the use of covariances or correlations as measures of association between variables, and depends upon the special mathematical properties of these measures. Problems of estimating communalities, determining the correct number of factors, rotating to simple structure, and naming the factors are not subjective, but, in principle, admit to only one correct solution. In contrast, nonmetric multidimensional scaling allows to use of ordinal measures of association, and the user is free to choose whatever measure best captures the interesting features of the data. The choice of the number of dimensions in which to represent the data and the orientation of the axes of the coordinate system are matters of informed judgment. There is no commitment to a "true" number of dimensions or a "true" coordinate system. In sum, the aim of factor analysis is to fit a specific model to the data; the aim of multidimensional scaling is to represent the data.

Special Problems With Dichotomous Item-level Data

Special problems are associated with factor analyzing dichotomous data. Specifically, dichotomous data cannot be modeled perfectly under the assumptions of factor analysis. That is, when variables can take on only two discrete values, they cannot be well described as the weighted sum of continuous factors. While there have been a variety of ad hoc solutions to this problem, the most defensible approach has been to apply the factor analytic model, not directly to the observed dichotomous variables, but to hypothesized unobservable continuous variables corresponding to each manifest response (Christopherson, 1975; Muthén, 1978). Though mathematically elegant, these models have been limited in their application to fairly small item sets (up to about 20 items) due to technical problems of estimation. Moreover, the models entail the additional assumptions that the factor scores and the hypothesized continuous variables have multivariate normal distributions.

In addition to problems arising from the dichotomous nature of item response data, individual test items are inherently less reliable than whole tests. Small fluctuations in an examinee's attention or performance during a test can have a big impact on the response to a single item, but are unlikely to significantly affect the total test score. This inherent unreliability is of concern in any item-level analyses, whether MDS, factor analysis, or some other technique. In order to obtain a stable solution, analyses of individual items must employ larger samples of examinees than analyses of test scores.

Method

Data

These analyses use data from the Reading subtest of the Metropolitan Achievement Test (MAT), Elementary Battery, Form F (Durost, Bixler, Prescott, Wrightstone, & Balow, 1970). This test, which is appropriate for assessing fourth-graders, consists of eight short pieces of text, each followed by four to eight 4-option multiple-choice questions. The questions are designed to

require: comprehending the literal meaning of information in the text; drawing inferences from the passage; identifying the best name or main idea of the passage; or determining the meaning of a word in context (Prescott, 1973). There are 45 items on the test.

Data were taken from the public-use tapes of the Morning Sample data from the Anchor Test Study (Loret, Seder, Bianchini & Vale, 1974). The tape contains test information for a nationally representative sample of approximately 63,000 fourth-grade children in over 400 schools. A systematic sample of every thirtieth record on the tape (N=2089) was used for these analyses. This sampling procedure assured proportionate representation of all the schools in the original sample because the tape was sorted by school.

Estimates of Similarity

Multidimensional scaling requires, as input, measures of association between all the objects to be scaled. Therefore, we needed an appropriate estimate of similarity between test items. The phi coefficient, which is the product-moment correlation between pairs of dichotomous variables, is often used as an index of association between test items. However, it suffers the serious limitation that its maximum possible value depends critically upon the item difficulties. The phi coefficient can reach a value of one, only if the two items being correlated are equally difficult. The upper limit on the phi coefficient is 0.82 for two items with difficulties as similar as 0.5 and 0.6; the upper limit drops to 0.65 if the difficulties are 0.5 and 0.7. Therefore, the use of phi coefficients in MDS would be likely to result in an artifactual dimension of difficulty.

Tetrachoric correlations were selected for use in this study because they are less sensitive to differences in item difficulties than phi coefficients and many other measures of association (Carroll, 1961). In general, the use of tetrachoric correlations has been challenged on several grounds. Some researchers have claimed that they are difficult to compute, that a sample matrix of tetrachoric correlations may not be Gramian, that they have large standard errors relative to product-moment correlations, and that they require hypothesizing unobservable continuous variables corresponding to each manifest binary variable. For our purposes none of these objections is sound. First, several efficient computational algorithms have been developed, so that calculating tetrachoric correlations is feasible with computing resources routinely available today. Second, the fact that a matrix of separately calculated sample tetrachoric correlations is sometimes non-Gramian is a technical problem of estimation, not a substantive problem. Algorithms might be constructed for constrained maximum likelihood estimation of positive semi-definite tetrachoric correlation matrices following a procedure similar to that of Bock and Petersen (1975). (The sample tetrachoric correlation matrix we calculated for these data was Gramian.) Third, regarding the standard error of tetrachoric correlations, the sample size used in this study (N=2089) was sufficient to presume adequate precision. Finally, the assumption of underlying continuous variables is irrelevant for nonmetric MDS. The technique requires only that the rank ordering of the similarities between variables be accurate. As long as this requirement is met, assumptions about underlying distributions of skills, whether accurate or not, are inconsequential.

We calculated the matrix of tetrachoric correlations using a special-purpose FORTRAN program. The algorithm used was due to Saunders and improved via Newtonian iteration, as described and implemented by Froemel (1971). The program calculated correlations based only on actual responses. That is, omitted responses were dropped from the analysis, rather than being scored as incorrect. This technique avoids computing correlations that are spuriously high because some examinees do not have time to complete the test.

Analyses

The matrix of tetrachoric correlations was scaled using the non-metric procedures of the KYST computer program (Kruskal, Young, & Seery, 1973). This program estimates the best final configuration of points from a given starting configuration by an iterative procedure designed to reduce stress (i.e., the mismatch between the rank orderings of the similarities and the calculated distances in the configuration), and then rotates the axes to principal components. The analysis was exploratory, and was conducted with no a priori notions about the number of dimensions that might be needed to represent the data. Therefore, we scaled the data in from one to six dimensions and looked at both the level of stress and the configuration of points in each solution to decide how to best represent the data. Initial analyses used Kruskal's stress formula 1 (SF1) and the Torsca starting procedure. Later, analyses used Kruskal's stress formula 2 and different starting configurations to see how the outcomes would compare, and as a check against the problem of local minima. The final representation of the data was selected based on stress information, visualizability, and the interpretability of the various solutions.

We sought patterns in the final configuration based on the following item characteristics: discrimination, difficulty, location of the item in the test, passage dependence, and item type. Item discrimination was measured by the point-biserial correlation between each item and the total test score. Item difficulty was defined as the proportion of examinees that answered an item correctly. Location refers to items classified as being at the beginning, middle or end of the test, depending on whether they are associated with the first three, middle two, or final three passages of text. Passage dependence was determined from results of a study by Tuinman (1973, personal communication, November, 1981) in which he gave the items from this test, but not the corresponding passages, to 1200 fourth-graders across the state of Indiana. We used, as the measure of passage dependence (pd), the proportion of Tuinman's examinees that could answer an item correctly without seeing the associated passage. Tuinman's sample, though large and broadly representative, is not strictly comparable to ours, so the value of pd may differ somewhat in our sample. However, for our analyses, this difference is unimportant as long as the rank order of pd values is similar in the two samples. Item type refers to one of four types of items that the publishers have identified in the test ("Content outlines," 1971): literal comprehension, inference, main idea, and vocabulary. Our judgment of item type (completed before the data were scaled) corresponds to the test publisher's judgment for all but five items, which the publisher labeled as measuring literal comprehension and which we think require some degree of inference. These five items will be distinguished when the results are presented.

We hypothesized that several other item characteristics (e.g., distractor similarity, syntactic complexity, and memory load) might be important to the

underlying test structure. However, the items did not vary systematically along these dimensions, and so we were unable to classify them reliably and unambiguously according to these features. Consequently, these characteristics were omitted from the present analyses.

Results and Interpretation

This section comprises two parts. The first part reports some general features of the data, describes how we selected a final MDS configuration, and suggests that a two-dimensional representation can reveal much of the underlying structure of data, even if more than two dimensions are apparent in the data. The second part describes and interprets patterns in the data, based on item characteristics.

Selection of a Representation

Different starting configurations can yield different MDS representations. One strives to find the configuration of points that best captures the underlying structure of the data. Our selection of a "best" representation was based on a low level of stress, interpretability of the configuration, and ease of visualization.

Stress. Figure 1 plots the minimum stress at each level of analysis by the number of dimensions. As one can see, the stress level falls at first, but then tapers smoothly as dimensions are added, giving little indication as to how many dimensions are needed to represent these data. Clearly, one dimension (SF1=0.313) is inadequate; at least two (SF1=0.203) and possibly three (SF1=0.159) are necessary. In more than three dimensions, stress decreases slowly, so it is unclear whether or not these dimensions might be meaningful. In this case, information other than stress level is particularly important for deciding how many dimensions are apparent in these data.

Interpretability. Analysis of item content revealed that the higher dimensions that are identified by rotating axes to principal components seem to be pulling out individual points or opposing sets of points that are unrelated to each other in terms of any features we could identify. These points do not seem to define dimensions in the data as a whole, but rather they seem to take advantage of the space created by additional dimensions to move away from the other points. This interpretation is supported by the fact that dimensions are not consistently defined by the same points as the representation moves into higher dimensional space. For example, in the three-dimensional solution, Dimension 3 is represented by Items 2, 17 and 44 at one extreme, and by Item 15 at the other extreme. When the solution goes into four dimensions, Dimension 3 is defined by Item 2 at one end and item 37 at the other. Items 15 and 44 have moved into Dimension 2 and Item 17 has moved into Dimension 4. As we include more dimensions, a pattern emerges: there are seven to nine points (about 17% of the items on the test) that tend to go off in their own directions, leaving the majority of points behind. We suspect, for two reasons, that this outcome is not simply due to error in the data but, rather, represents true idiosyncracies in these items. First, the sample is large and nationally representative. Thus, we expect that if this experiment were repeated most of the same points would appear as outliers. Second, in reading the items, prior to scaling, we hypothesized that several of these outlying items measure something other than (or in addition to) reading comprehension.

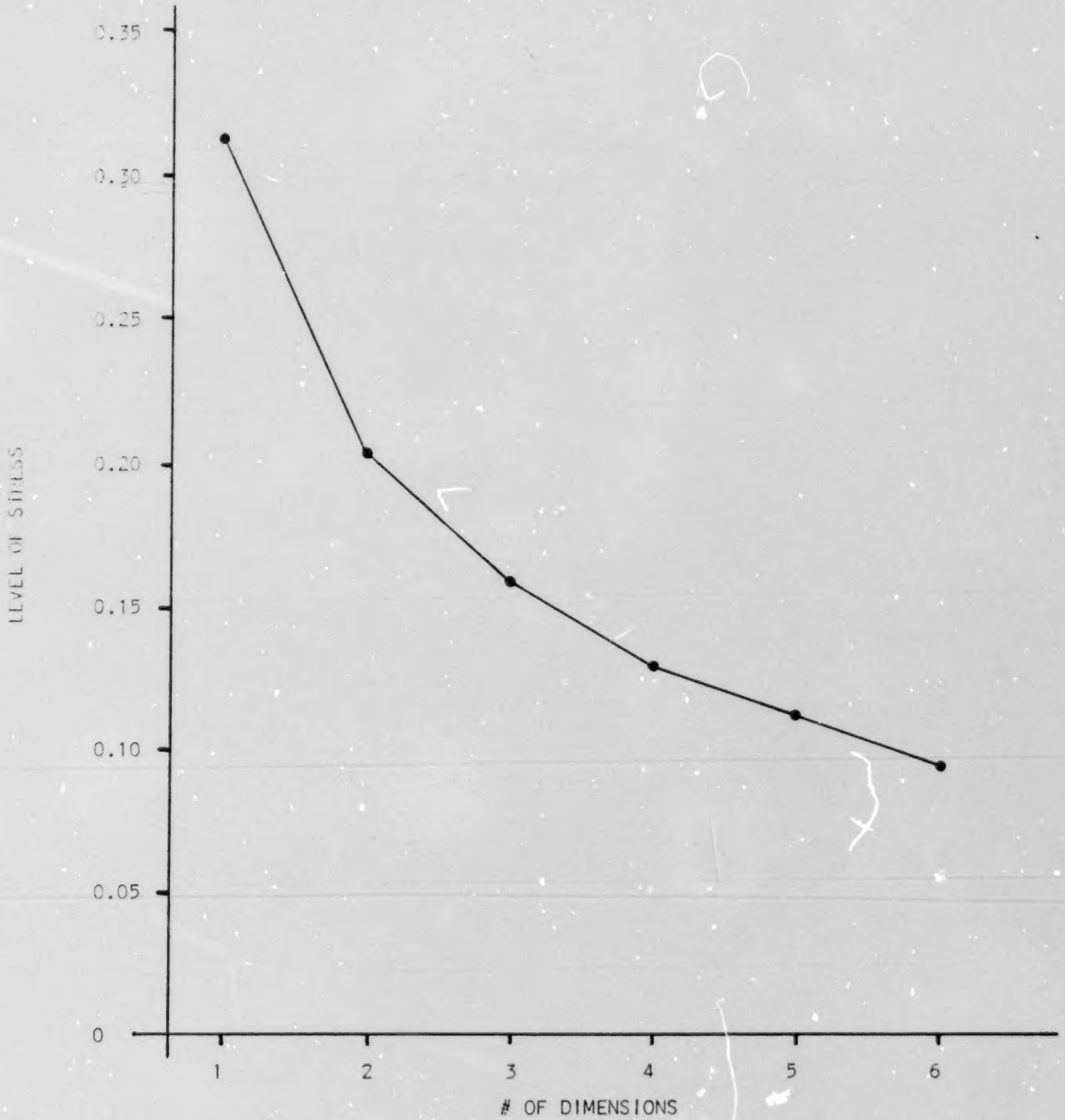

Figure 1. Plot of stress (Kruskal's stress formula 1) by number of dimensions in multidimensional scaling. (Data from Reading subtest of Metropolitan Achievement Tests, Elementary Battery, Form F.)

Visualizability. As described above, the axes identified by rotating the cofiguration to principal components seem to be defined by idiosyncratic outlying items. Therefore, one would not necessarily expect them to be psychologically meaningful. Indeed, apart from Dimension 1, the principal axes identified do not seem to capture the psychologically important dimensions in these data. This does not matter in two dimensions, because one may draw axes wherever they seem appropriate. However, in three or more dimensions, the axes chosen become crucial: if they are not the psychologically interesting ones, it is very difficult to visualize the data so that the important dimensions can be identified. Therefore, we chose to present the data in two dimensions, expecting that the major trends will be apparent--though the dimensions may not appear orthogonal to each other--and that this representation is less likely than higher dimensional ones to hide important patterns. To no surprise, the seven to nine points that tend to go off in their own dimensions move toward the periphery of the two-dimensional representation.

Interpretability again. Using various starting configurations, we found three constellations that fit the data equally well in two dimensions (for each, SF1=0.203). As would be expected, these solutions are quite similar, varying in the position of only a few points. We had one dilemma in selecting which configuration to present here. There is one item that moves a substantial distance between solutions, appearing in two distinct regions. As one would hope, both of these locations make sense in terms of the item's characteristics. Ultimately, we chose the configuration in which this item seems to be located according to its type. However, we will describe both substantive interpretations for this peripatetic item when we discuss the patterns in the data, below.

Patterns Based on Item Characteristics

Figures 2 and 3 depict the selected two-dimensional representation of the data with items identified according to different characteristics. These characteristics reveal several strong trends.

Item discrimination. One expects to see a clear relationship between item discrimination and the MDS representation because both indicate how similar each item is to all other items simultaneously. Item discrimination, as summarized by the point-biserial correlation, estimates the degree of association between each item and the composite of all items on the test. MDS locates objects in space according to their similarity with each other object in the space. Thus, though these summaries are computed in very different ways, they tell similar stories. Specifically, the most highly discriminating items (i.e., those that share most in common with the other items) should cluster in the center of the MDS representation, and the least discriminating (i.e., the most idiosyncratic) items should fall around the periphery. Items of intermediate discrimination should fall along a gradient in between. Figure 2 (the MDS configuration with points coded by degree of item discrimination) does, indeed, exhibit this pattern, and contributes evidence that this MDS representation accurately reflects the structure of the data.

Item difficulty. Figure 3a (with points coded according to the proportion of examinees answering correctly) shows a distinct trend according to item difficulty. The easiest items ($p \geq 0.80$) cluster together, and

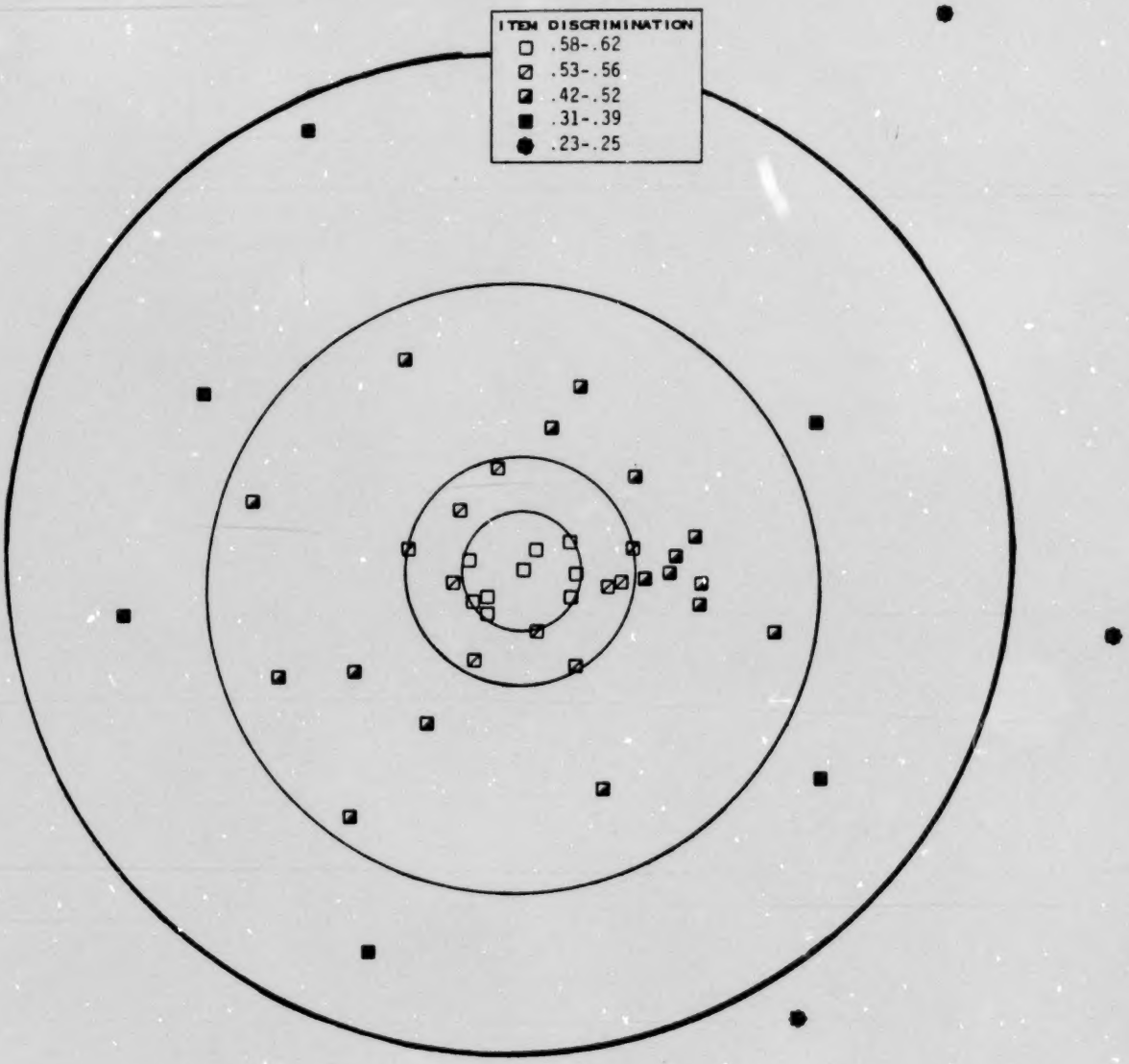

Figure 2. Multidimensional scaling of 45 items from the Reading subtest of the Metropolitan Achievement Tests, with items identified according to level of discrimination (point-biserial correlation of item with total test score).

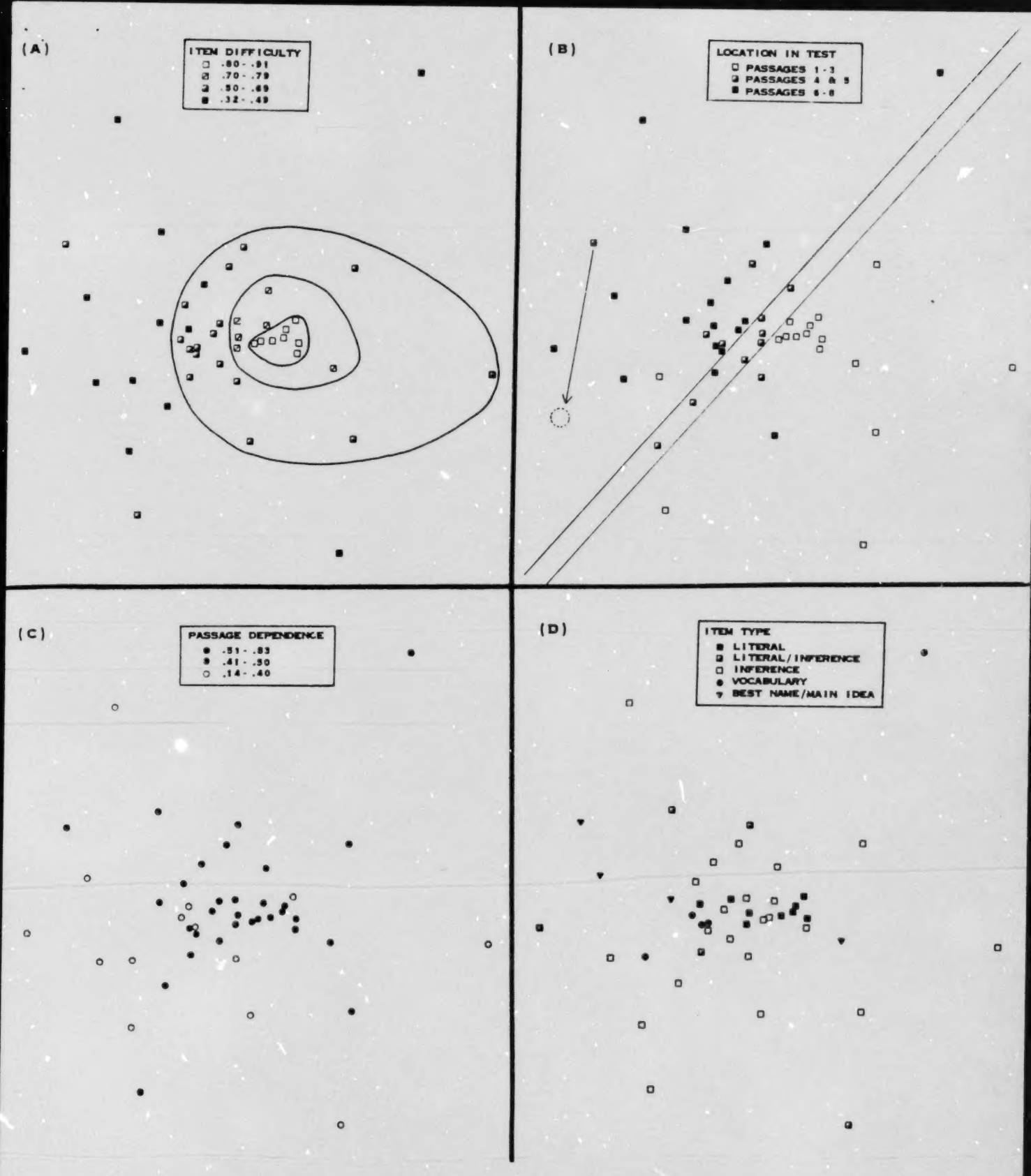

Figure 3. Multidimensional scaling of 45 items from the Reading subtest of the Metropolitan Achievement Tests, with items identified according to (a) difficulty (proportion correct), (b) location in test, (c) passage dependence (proportion correct without passage), and (d) item type. Arrow in (b) indicates alternative location for item.

progressively harder items fall into nested egg-shaped rings around them. Notice that the items are not evenly distributed inside the rings; rather, they are more densely packed on the left. Indeed, if one imagines a vertical line drawn at the right edge of the cluster of easiest items, the points to the left of it fan out along the dimension of item difficulty, and only six widely scattered points fall to the right of that vertical line. The obvious question at this point is: what features distinguish the items on the left and right, that is, why don't the six dispersed points fall in with the rest? Apparently, these items each measure something idiosyncratic. Indeed, the three most widely scattered ones have the lowest point-biserial correlations of all items on the test, and two more have quite low point-biserials (compare Figure 2). Not surprisingly, these five points move into higher dimensions if allowed to.

Location in the test. The data also show a strong pattern according to the location of the items in the test (Figure 3b). Items from the beginning of the test (those associated with the first three passages) generally fall into the lower right portion of the configuration; items from the end (those associated with the final three passages) fall in the upper left; and items from the middle two passages generally fall along a diagonal line separating the early and late items. Notice the arrow in Figure 3b that indicates the two locations of the "peripatetic" item mentioned in the section on selecting a representation. This item, which is from the middle of the test, can be situated closer to the other middle items without altering the level of stress in the total configuration.

It is not surprising that Figures 3a and 3b (item difficulty and location) show a degree of correspondence, because the test is designed to have easier questions at the beginning and harder questions at the end. However, the correspondence is not perfect; some items near the beginning of the test are more difficult, and these appear on the opposite side of the configuration than the items that are more difficult from the end of the test. Also, the difficult early items spread apart from each other, indicating that they may be measuring idiosyncratic skills or knowledge. This interpretation is, again, supported by the fact that these peripheral items move out into their own space if more dimensions are added to the analysis. The difficult items from the end of the test, with a few exceptions, are more densely packed. This pattern suggests that most of these items measure relatively similar knowledge or skills. However, three of the last items seem to measure idiosyncratic skills--they sit around the periphery, and tend to move into their own higher-dimensional space when it is available.

Passage dependence. Figure 3c shows the configuration with items identified according to passage dependence, i.e., Tuinman's estimate of the proportion of examinees who can answer items correctly without reading the passage (pd). The pattern here is striking. Most of the items that can be answered easily without reading the passages--that is, items which over half of Tuinman's sample answered correctly without the passages available--(solid circles in Figure 3c) cluster in the middle of the configuration. Indeed, the central points consist almost exclusively of this type of item. Of the most passage dependent items--items that 40% or fewer of Tuinman's examinees could answer correctly without reading the passage--(open circles in the figure) only four fall in with this group, and their presence might be explained. Of these four items, two are vocabulary items that may require the presence of the passage, but do not require a thorough reading of it. It is possible to

answer these items simply by locating the bold-faced word in the passage and comprehending its meaning within a single phrase or sentence. The other two are literal comprehension items that can be answered by reading only the final sentence in the passage. (In one case, this sentence immediately precedes the question.)

It seems that the MDS representation centers around a large and relatively dense cluster of items that might be answered with little or no reference to the passage with which they are associated. Content analyses of these items suggests that half of them might be answered using general knowledge (e.g., that many diseases are caused by germs, not scientists, penicillin or medicine), and the other half probably require some aspect of test-wiseness (e.g., gleaning information from other questions, or second-guessing the test writer to choose the most plausible sounding option).

Item type. Figure 3d plots the MDS configuration with items identified by type. (The items that the publisher labeled as measuring literal comprehension and which we think require some degree of inference are identified in the figure as literal/inference.) Inference items, which are most numerous, scatter about the configuration. All of the items that we judged as requiring literal comprehension fall within the center of the configuration. The five items on which we disagreed with the test publisher are scattered about the figure, one of them lying close to the center. The test contains five vocabulary items that require the examinee to identify the meaning of a word in the passage (represented in Figure 3d as circled dots). Three of these items cluster tightly together in the center of the solution. A fourth lies nearby (it is almost as close to the other vocabulary items as it is to anything), and the last one sits off by itself in the upper right corner. This final item is the most unusual item on the test: it has the lowest point-biserial correlation and moves out furthest when new dimensions of space are added. Finally, the test contains four items that ask for the best name or the main idea of the passage (solid triangles in the figure). Three of these items radiate out in a line to the left of the center. The fourth sits on the opposite side of the large central cluster of points.

In sum, item type seems to explain less of what is going on in the data than any of the other features identified, partly because there are so few items designed to measure anything other than literal comprehension or inference ability. The main idea and vocabulary items may measure two specific skills: Three of the four main idea items seem to relate more closely to each other than they do to other items. Three of the five vocabulary items form a densely packed group, indicating that examinees who answer one of them correctly also tend to pass the other two. (This result is not necessarily expected, because vocabulary items measure understanding of distinct words, and knowledge of one word need not imply knowledge of another.) A fourth vocabulary item seems to be somewhat related to the other three, but it is somewhat more difficult. The fifth appears to be quite different from anything else on the test. It is less clear what the literal and inference items measure. The items that we identified as requiring literal comprehension all fall within a central group of points that seem to require little or no reading of the passage (compare Figure 3c). Inference items and the other "literal" items scatter about the MDS representation and apparently measure individual skills or knowledge.

Discussion

Multidimensional scaling of these reading test data reveals some surprising information about the underlying structure of the test. The test centers around a stable set of items that apparently can be answered with little or no reference to the associated passages. This result is surprising and makes sense at the same time. It is surprising because one does not expect a test of "reading skill" to contain so many items for which reading the text is not required. The configuration makes sense because the items in the center should have most in common with the entire set of items, and therefore ought to require some general skills or knowledge. In this case, the general abilities seem to be possession of some specific common knowledge, reading and comprehension of multiple-choice test items, and perhaps a bit of test-wiseness.

Most of the more difficult items scatter outside the central group. It appears that slightly different specific abilities are required to solve each of these items. This result is consistent with an original prediction made by Guttman for the structure of mental ability test batteries. Guttman (1954) described a hypothetical radex structure, with simpler tests in the center and progressively more complex tests radiating out. It is significant that Guttman (1965), and more recently Marshalek, Lohman, and Snow (1983), did not find this result when they scaled batteries of mental tests. Rather, they found that more complex (Guttman called them "rule-inferring") tests fell in the middle and simpler (or "rule-applying") tests fell near the periphery. The apparent contradiction between these results and ours may be a matter of word usage. Our results are entirely consistent with those of Guttman and Marshalek et al. if one thinks in terms of a continuum of generality, that is, with tests (or items) that require general abilities in the middle and tests that require specific abilities radiating toward the outside.

A facet of item difficulty that corresponds somewhat to item location in the test also seems apparent in these data. The fact that the two facets correspond imperfectly indicates that the determinants of item difficulty are different for earlier and later items. Specifically, the early difficult items seem to measure idiosyncratic skills or knowledge. Most of the later ones seem to measure skills that probably involve high levels of vocabulary, syntactic complexity, passage length and abstractness of ideas. A few of the later difficult items may measure idiosyncratic skills or knowledge.

Information about item type reveals least about the test since 80% of the items are designed to measure either literal comprehension or ability to make inferences. Most of the literal comprehension items fall in the middle of the representation, and so seem to measure a general ability on this test. We do not know whether they actually measure literal comprehension ability or general knowledge and test-wiseness because most of these items can be answered by many examinees without reading the passage. The inference items spread apart from each other, and do not seem to measure a unitary skill.

Summary

This paper makes three primary contributions to the study of construct validity. First, it proposes that nonmetric multidimensional scaling, with its limited assumptions, is a useful technique for representing the interrelationships between items in a test. Second, it suggests that typical

problems associated with scaling dichotomous variables can be avoided by using tetrachoric correlations as input to the multidimensional scaling. Finally, it demonstrates the utility of the suggested procedures by applying them to actual data from the Reading subtest of the Metropolitan Achievement Tests.

References

Bock, R. D., & Petersen, A. C. (1975). A multivariate correction for attentuation. Biometrika, 62, 673-678.

Carroll, J. B. (1961). The nature of the data, or how to choose a correlation coefficient. Psychometrika, 26 (4), 347-372.

Christopherson, A. (1975). Factor analysis of dichotomized variables. Psychometrika, 40, 5-32.

Content outlines (Metropolitan Achievement Tests Special Report, 1970 Edition, Report No. 2). (1971). New York: Harcourt, Brace, Jovanovich.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? In W. B. Schrader (Ed.), Measuring achievement: Progress over a decade (New Directions for Testing and Measurement, No. 5, pp. 99-108). San Francisco: Jossey-Bass.

Durost, W. N., Bixler, H. H., Wrightstone, J. W., Prescott, G. A., & Balow, I. H. (1970). Metropolitan Achievement Tests, Form F (Elementary Level). New York: Harcourt, Brace, Jovanovich.

Froemel, E. C. (1971). A comparison of computer routines for the calculation of the tetrachoric correlation coefficient. Psychometrika, 36, 165-174.

Guttman, L. (1954). A new approach to factor analysis: The radex. in P. F. Lazerfeld (Ed.), Mathematical thinking in the social sciences. Glencoe, IL: Free Press.

Guttman, L. (1965). The structure of interrelations among intelligence tests. In Proceedings of the 1964 invitational conference on testing problems. Princeton, NJ: Educational Testing Service.

Hathaway, W. E. (Ed.). (1983, September). Testing in the schools (New Directions for Testing and Measurement, No. 19). San Francisco: Jossey-Bass.

Herman, J. L., & Dorr-Bremme, D. W. (1983, September). Uses of testing in the schools: A national profile. In W. E. Hathaway (Ed.), Testing in the schools (New Directions for Testing and Measurement, No. 19, pp. 7-17). San Francisco: Jossey-Bass.

Kruskal, J. B. & Wish, M. (1978). Multidimensional scaling. Beverly Hills: Sage.

Kruskal, J. B., Young, F. W., & Seery, J. B. (1973). How to use KYST, a very flexible program to do multidimensional scaling and unfolding. Unpublished paper, Bell Telephone Laboratories.

Loret, P. G., Seder, A., Bianchini, J. C., & Vale, C. A. (1974). Anchor test study--The equating and norming of selected reading achievement tests (grades 4, 5, and 6). Washington, D. C.: U. S. Department of Health, Education and Welfare, Office of Education.

Marshalek, B., Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. Intelligence, 7, 107-127.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. Psychometrika, 43, 551-560.

National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform. Washington, D. C.: U. S. Government Printing Office.

Prescott, G. A. (1973). Manual for interpreting--Metropolitan Achievement Tests. New York: Harcourt, Brace, Jovanovich.

Tuinman, J. J. (1973). Determining the passage dependency of comprehension questions in five major tests. Reading Research Quarterly, 9, 206-223.